Reproducibility issues in science, is P value really the only answer?

Jean Gaudart, Laetitia L. Huiart, P. J. Milligan, Rodolphe Thiebaut, Roch Giorgi

LETTER

Reproducibility issues in science, is *P* value really the only answer?

Johnson describes the lack of reproducibility of scientific studies, attributed, according to the author, to the low level of significance (1). We appreciate the quality of this work and its importance for the interpretation of statistical evidence. These results should be considered in statistical guidelines. Nevertheless, we would like to point out some important points not thoroughly discussed in this publication.

Not publishing "nonsignificant" results leads to the well-known publication bias whereby studies with low statistical power are underrepresented. This bias would become more severe, despite recommendations to allow for publication of "negative" results. Lowering the significance level will further increase the type II error, which is clinically as important as type I error. Focusing only on the type I error may lead to an excessive false nondiscovery rate. In the case of severe diseases, it is not uncommon to fix a significance level at 0.1 (2), at the early stages, to avoid excluding an effective treatment. Johnson argues that this may be corrected by increasing the sample size. However, increasing the size of clinical trials will reduce their feasibility and increase their duration. Aside from these issues, including more patients means exposing more patients to an experimental treatment and may challenge the equipoise concept.

The issue of fixing a threshold defining significance refers to the Fisher–Pearson controversy. Estimating a *P* value is needed to quantify the strength of evidence. However, fixing a threshold is needed to make a decision controlling for the risk of type I and type II error. Actually, regarding the issue addressed by Johnson, it would be interesting to assess if a priori specification of the threshold is required, or if research results could be compared using the *P* value and the magnitude of the tested statistic.

The issue of significance level is only the tip of the iceberg. Indeed, design issues should not be overlooked when discussing lack of reproducibility. Selection bias leads to extrapolation of results to a population different from the target population (3). Furthermore, the "poor reporting" practice highlighted by Altman et al. (4) and the lack of compliance to reporting recommendations (e.g., Consolidated Standards of Reporting Trials) hinder a proper assessment of the quality of the study and hide selection bias or misuse of statistical tests; the latter leads to nonreproducibility of the reported research. In an extreme example, monthly American Air passengers and the Australian electricity production in the late 1950s are highly correlated (Pearson's correlation = 0.88, $P = 8.8 \times 10^{-13}$) without any meaning.

The causality criteria defined by Hill (5) highlight other important considerations in the interpretation of results. Reliance on *P* values remains surprisingly widespread, but good decision making depends on the magnitude of effects, the plausibility of scientific explanations of the mechanism, and the reproducibility of the findings by others.

Jean Gaudart[a,1], Laetitia Huiart[b], Paul J. Milligan[c], Rodolphe Thiebaut[d], and Roch Giorgi[a]

[a]Aix-Marseille Université, Institut National de la Santé et de la Recherche Médicale, Unité Mixte de Recherche 912, 13005 Marseille, France; [b]Clinical Epidemiology Department, Centre Hospitalier Universitaire La Réunion, 97405 Saint-Denis, France; [c]London School of Hygiene and Tropical Medicine, London WC1E 7HT, United Kingdom; and [d]Université de Bordeaux, Institut National de la Santé et de la Recherche Médicale, Research Centre Epidemiology and Biostatistics, Institut de Santé Publique, d'Epidémiologie et de Développement, 33076 Bordeaux, France

1 Johnson VE (2013) Revised standards for statistical evidence. *Proc Natl Acad Sci USA* 110(48):19313–19317.
2 Korn EL, et al. (2001) Clinical trial designs for cytostatic agents: Are new approaches needed? *J Clin Oncol* 19(1):265–272.
3 Horton R (2000) Common sense and figures: The rhetoric of validity in medicine (Bradford Hill Memorial Lecture 1999). *Stat Med* 19(23):3149–3164.
4 Altman DG, Moher D, Schulz KF (2012) Improving the reporting of randomised trials: The CONSORT Statement and beyond. *Stat Med* 31(25):2985–2997.
5 Hill AB (1965) The environment and disease: Association or causation? *Proc R Soc Med* 58:295–300.

Author contributions: J.G., L.H., P.J.M., R.T., and R.G. wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. E-mail: jean.gaudart@univ-amu.fr.